DOCUMENT RESUME

ED 078 002 TM 002 814

AUTHOR Hoover, H. D.; Plake, Barbara

TITLE An Empirical Comparison of Selected Two-Sample Hypothesis Testing Procedures Which Are Locally Most Powerful Under Certain Conditions.

PUB DATE 73

NOTE 14p.; Paper presented at American Educational Research Association Meeting (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29

DESCRIPTORS *Comparative Analysis; *Hypothesis Testing; Models; Probability; *Research Methodology; *Statistical Analysis; Tables (Data); Technical Reports

ABSTRACT

The relative power of the Mann-Whitney statistic, the t-statistic, the median test, a test based on exceedances (A,B), and two special cases of (A,B) the Tukey quick test and the revised Tukey quick test, was investigated via a Monte Carlo experiment. These procedures were compared across four population probability models: uniform, beta, normal, and double exponential. Sample sizes of (5.5), (10,10), (20,20), (5,10), and (5,20) were among those used. Results indicate the median test should be considered for distributions which certain outliers. The exceedances tests can be powerful alternatives to more standard procedures if the underlying distributions are platykurtic. (Author)

ED 078002

TM 002 814

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

AN EMPIRICAL COMPARISON OF SELECTED TWO-SAMPLE HYPOTHESIS TESTING PROCEDURES WHICH ARE LOCALLY MOST POWERFUL UNDER CERTAIN CONDITIONS

H. D. Hoover, University of Iowa
Barbara Plake, James Ford Bell Technical Center

INTRODUCTION

In the last few years, a great deal of information has been published regarding the robustness of the t-statistic and other normal distribution theory hypothesis testing procedures. In general, these procedures are remarkedly robust when the underlying assumptions are violated, especially with respect to control over errors of the first type. Exceptions occur when both the variances and sample sizes are unequal and under some conditions of rather extreme non-normality, primarily skewness. A very comprehensive review of the research on the robustness of the Student-procedure is reported by Hatch and Posten (1966).

While a great deal of research has been conducted on the robustness of the t-statistic, and a few of its distribution free competitors, this research has tended to focus on a rather narrow definition of robustness; i.e., the control over Type I errors. Violation of the assumptions necessary for the exactness of any hypothesis testing procedure also affects its control over Type II errors. Conditions of non-normality and variance heterogeneity, while not always detrimental to the performance of the t-statistics control over the nominal significance level, sometimes have a very noticeable effect on the t-tests power, especially relative to other hypothesis testing procedures, Pratoomraj (1970).

AERA New Orleans '73

FILMED FROM BEST AVAILABLE COPY

Purpose of the Study

As is well known, many of the distributions which exist in educational and psychological research are non-normal in nature. Many carefully constructed standardized tests yield raw score distributions which are by necessity bounded and relatively flat or platykurtic in nature. Some of these distributions are, in fact, nearly rectangular or uniform, Brandenburg (1972). Another, contrasting, situation is one in which an occasional large measurement error will produce a highly disparate observation. This tends to create underlying distributions which are leptokurtic (peaked).

The major objective of this study was to investigate the relative power of four two-sample hypotheses testing procedures across four different underlying distributions for five variations of sample size. The four statistical procedures investigated were (1) the t-test (t), (2) the Mann-Whitney U test (U), (3) the median test (r), and (4) two variations of a test based on exceedances: a procedure described by Hajek (1969) which will be designated by (A,B) and a procedure recommended by Tukey (1959) referred to as (A+B).

Description of Statistics Investigated and Probability Models Sampled

In order to empirically determine the relative Type I error control and power of the various hypotheses testing procedures, four probability distributions were used as sampling models. Each of these distributions was continuous and symmetric, but each differed primarily in tail weight or degree of kurtosis = $K = E[(X - \mu)^4]/\sigma^4$. These distributions were: 1) the double exponential, 2) the normal, 3) the uniform, and 4) a lambda distribution (Ramberg and Schmeiser, 1972) with tail weight ($K = 2.3$) between the normal and uniform distributions.

Among the rank tests the two-sample median test is locally most powerful when the underlying distributions are double exponential. The double exponential distribution is characterized by its long (heavy) tails (K = 6.0). The Mann-Whitney U test is the uniformly most powerful rank test when the underlying distributions are logistic. The logistic distribution is somewhat lighter (or shorter) tailed (K = 4.2) than the double exponential, but still heavier tailed than the normal probability model (K = 3.0). It is well known that the t-statistic is uniformly most powerful if the underlying distributions are normal. The two tests based on exceedances (A,B) and (A+B) are each locally most powerful, for different alternatives, when the sample distributions are uniform (K = 1.8).

The test statistic for the (A,B) test used in testing H_o: F(x) = F(y), against H_1: F(x) > F(y), is the ordered pair (a,b) where a is the number of y's greater than the largest x, and b is the number of x's less than the smallest y. The (A,B) test assumes that the pairs are ordered by the following rule:

$$(A,B) > (A',B') \text{ if } \begin{cases} \text{either } \min(A,B) > \min(A',B') \\ \text{or } \min(A,B) = \min(A',B') \text{ and } (A+B) > (A'+B') \end{cases}$$

$$(A,B) = (A',B') \text{ if } \begin{cases} \text{either } A' = A,\ B' = B \\ \text{or } A' = B,\ B' = A \end{cases}$$

Then the pair (A,B) whose values (a,b) are ordered as above provides a one ended test of H_o: F(x) = F(y). The (A,B) test is locally most powerful for uniform distributions with small mean differences.

The test statistic for the (A+B) test is a + b where a and b are the same as for the (A,B) test. This procedure is locally most powerful for uniform distributions with "large" mean differences.

Graphs and density functions of the probability distributions sampled are given in Figure 1. Since these graphs do not clearly show the distinction in tail weights of the distributions, each distribution was rescaled to have the same median and .95 quantile as the standardized normal. This comparison of the tails of the four distributions is shown in Figure 2.

Procedures

The procedure used to generate the empirical sampling distributions of the hypothesis testing procedures investigated is described in the following steps:

1. Vectors of m + n elements, randomly drawn from each of the four population distributions, were obtained. The first m elements from an X-universe having mean μ_x and variance σ_x^2 and the remaining n from a Y-universe with mean μ_y and variance σ_y^2.

   Each of the m + n elements in the vector was obtained by generating a uniform random number between zero and one, which was regarded as a relative cumulative frequency of the uniform distribution. The random variable for each of the other distributions investigated (lambda, normal, double exponential) was then obtained through what amounted to an area transformation.

2. Five combinations of sample size (m,n)[(5,5); (10,10); (20,20); (5,10); (5,20)] and five values of Δ (0(1)4) were selected for investigation for each of the four probability models.
   $\Delta = (\mu_x - \mu_y)(\sigma_x^2/m + \sigma_y^2/n)^{-1/2}$

3. For each vector of m + n observations, the statistics t, U, r, (a,b), and (a+b) were computed. This procedure was repeated 1000 times for each combination of (m,n), population distribution, and Δ-value.

4. For each of the above replications the test statistics were referred to their respective .05 two-ended critical values. Since critical values corresponding to a significance level of exactly .05 do not ordinarily exist for the rank type procedures, a randomization process was used which insured that each would have a nominal level of .05.

Results

The empirical Type I error and power values (times 1000) obtained for this investigation are presented in Table 1. In general, these results are consistent with predictions obtained from asymptotic theory. In the discussion which follows the various hypotheses testing procedures will be compared for the different population models sampled. Of the two tests based on exceedances all references will be to (A+B). Little practical difference existed between (A+B) and (A,B) and because of the simpler decision rule associated with A + B it seems to be the preferable procedure.

The results may be summarized by sampled distribution as follows:

1. Double exponential. Across the various sample sizes studied both t and U exhibit excellent power. There appears to be little reason to prefer either of these procedures although U was slightly more powerful for the larger equal sized samples. The most surprising result for this population model was the very poor performance of the median test (r). While this procedure is the locally most powerful (Δ small) of the rank tests for double exponential distributions, the only case in which it was in any way comparable to t and U was when $m = n = 20$. Considering the manner in which the (A+B) procedure is defined it performed surprisingly well except for $m = n = 20$.

2. Normal. As was expected, t was the superior procedure for this case. However, as is well known, the Mann-Whitney statistic performs very well when the underlying distributions are normal. Once again (r) was inferior to (A+B) except when $m = n = 20$.

3. Lambda. For this relatively flat population model (K = 2.3), t was superior to all statistics investigated. There appears to be little reason to prefer either U or (A+B) and both would seem to be reasonable alternatives to the t-statistic. The median test is noticeably less powerful than any other across all sample size combinations.

4. Uniform. (A+B) and t are the preferable procedures for distributions of this type. The t-statistic is slightly more powerful than (A+B) in the m = n = 5 case and there appears to be little difference between the two methods for $m \neq n$. For the larger equal sample sizes (A+B) is the superior method, markedly so in m = n = 20 case. Although less powerful than t and (A+B), the U statistic performs reasonably well for rectangular distribution types. This is especially true relative to r which is markedly inferior to all procedures.

Selected results from Table 1 discussed above are illustrated in Figures 3 through 7.

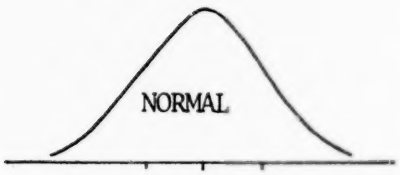
In summary, it appears that t is probably overall the superior statistic although for "heavy" tailed distributions U is a very competitive alternative and for "lighter" tailed underlying densities the tests based on exceedances are attractive alternatives, especially (A+B) because of its simplicity. With the exception of large samples from leptokurtic population models the median test has little to offer relative to the other procedures investigated.

TABLE 1

Empirical Power Functions for Each of Five Test Statistics (T') When Sampling from Selected Probability Models

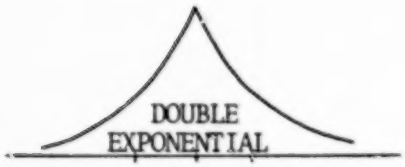
$(\Delta = (\mu_y - \mu_x)(\sigma_x^2/m + \sigma_y^2/n)^{-1/2}, \alpha = .05)$

MODEL		DOUBLE EXPONENTIAL					NORMAL					LAMBDA					UNIFORM				
m,n	T' \ Δ	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
	t	44	150	479	792	918	48	125	419	773	941	52	122	398	765	947	54	125	385	795	960
	U	44	170	465	751	875	46	128	399	741	916	43	118	378	731	934	47	110	350	710	940
	r	61	131	321	579	717	45	96	224	470	683	57	80	205	448	658	53	69	170	388	626
m=5	A,B	45	158	436	720	862	45	124	366	714	894	45	111	348	730	927	45	106	359	731	940
n=5	A+B	48	165	465	748	871	45	133	387	742	915	46	117	359	735	926	44	117	358	716	944
	t	52	169	493	822	950	56	163	455	832	970	54	165	443	834	972	55	163	437	829	977
	U	48	183	527	844	958	47	147	405	786	956	48	132	372	760	950	48	139	374	735	933
	r	51	184	498	774	928	50	132	290	570	827	55	104	255	507	761	54	99	195	390	645
m=10	A,B	55	114	316	567	733	55	119	357	666	876	55	124	413	771	945	55	190	575	899	981
n=10	A+B	53	134	363	630	802	51	142	397	705	898	53	141	412	783	950	50	188	559	892	979
	t	31	183	540	823	958	34	178	504	826	967	37	176	491	825	968	41	174	496	818	974
	U	40	228	634	891	983	40	173	462	801	960	40	155	434	781	946	40	160	443	759	939
	r	40	224	557	856	959	37	130	309	609	839	40	95	245	522	765	39	84	200	404	629
m=20	A,B	49	75	119	310	509	49	97	291	515	736	49	127	410	733	930	49	329	830	969	998
n=20	A+B	42	85	237	373	584	43	105	335	572	800	43	132	446	758	944	44	313	814	972	1000
	t	44	187	483	799	943	45	169	434	787	959	47	157	433	789	963	49	157	427	787	967
	U	36	202	497	811	930	35	151	420	762	943	41	153	391	738	939	40	154	390	707	935
	r	56	153	384	656	799	38	112	271	470	725	43	81	220	423	709	47	90	192	355	673
m=10	A,B	51	136	356	624	802	51	127	362	674	903	58	142	379	728	945	48	182	465	798	958
n=5	A+B	49	160	402	680	849	50	138	376	722	909	54	136	374	730	937	51	155	419	774	954
	t	36	191	514	812	964	37	178	472	809	972	45	175	460	802	975	46	173	442	803	980
	U	38	202	538	807	941	38	159	412	751	958	38	151	372	727	957	38	149	366	690	938
	r	52	204	482	676	840	53	132	332	571	812	55	114	281	523	791	48	86	221	466	782
m=20	A,B	50	100	238	433	662	50	139	292	542	795	50	171	364	632	875	50	227	540	765	931
n=5	A+B	44	157	309	546	747	43	161	350	652	885	46	173	405	726	954	46	209	515	839	989



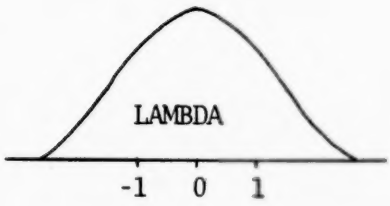
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-1/2x^2}$$

$-\infty < x < \infty$



$$f(x) = \frac{1}{2} e^{-|x|}$$

$-\infty < x < \infty$

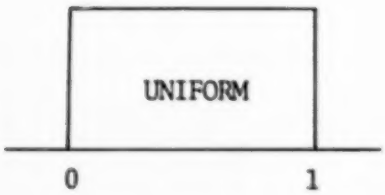


(inverse of the CDF)

$$x = F^{-1} = \frac{u^{.35} - (1-\mu)^{.35}}{.385}$$

where μ is uniform $0 \leq \mu \leq 1$

$|x| < 2.6$



$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

FIGURE 1

PROBABILITY DISTRIBUTIONS SAMPLED

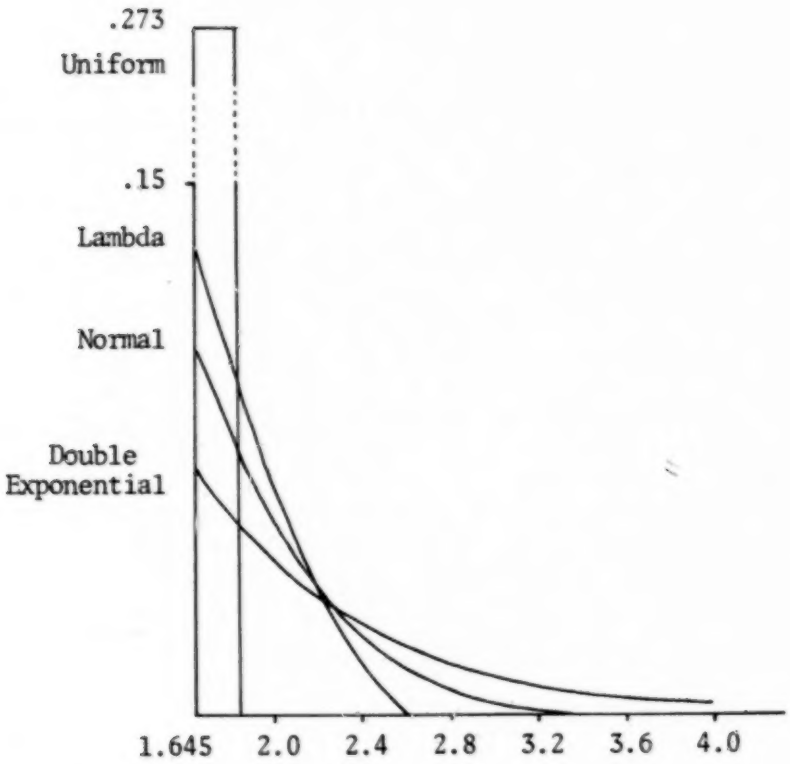


FIGURE 2

UPPER 5% TAILS OF DISTRIBUTIONS SAMPLED

FIGURE 3

Empirical Power Values and Smoothed Power Curves
for t, U, r, and A+B for Double Exponential Distributions
$m = n = 10$, $\alpha = .05$

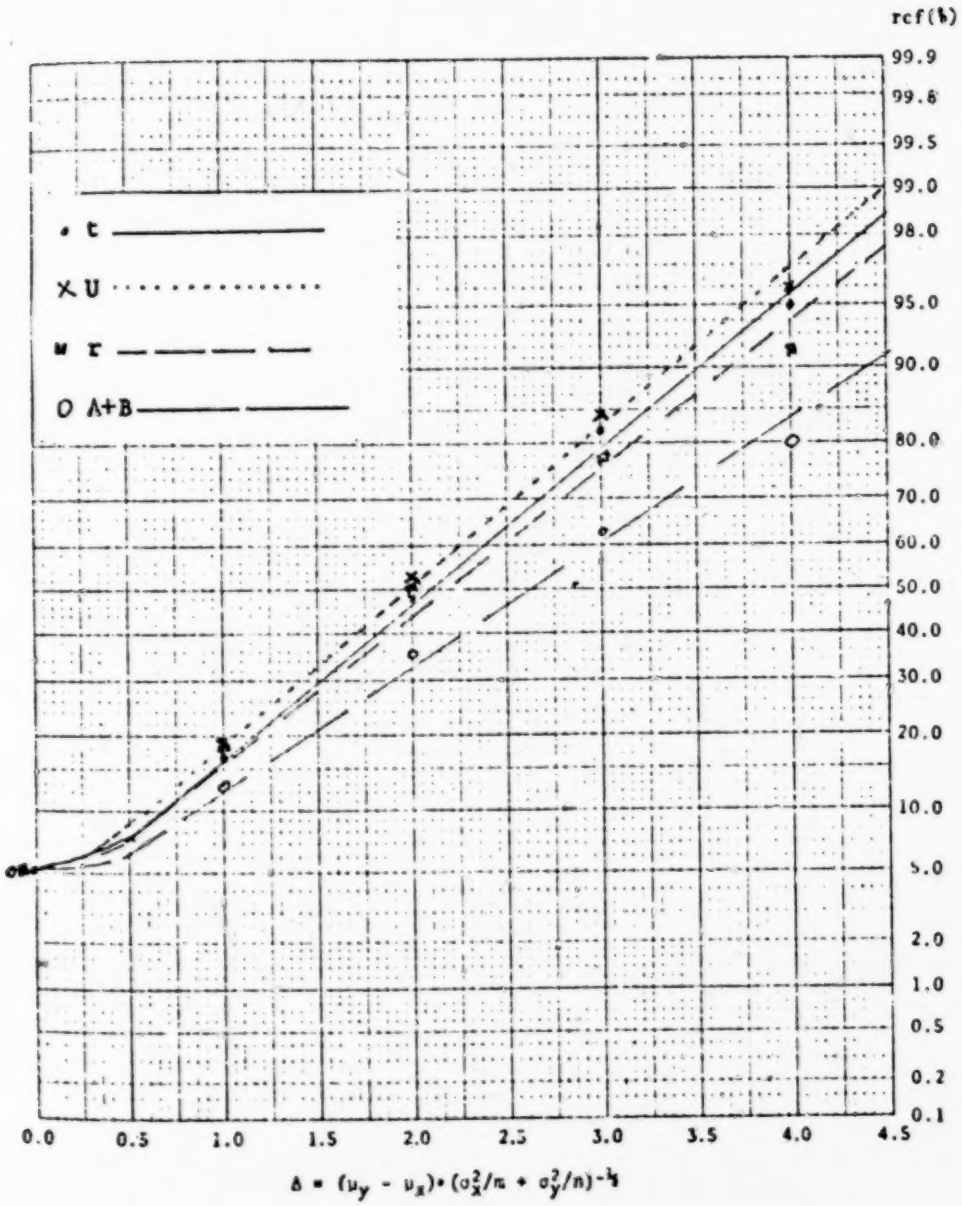


FIGURE 4

Empirical Power Values and Smoothed Power Curves
for t, r, U, and A+B for Lambda Distributions
m = n = 20, α = .05

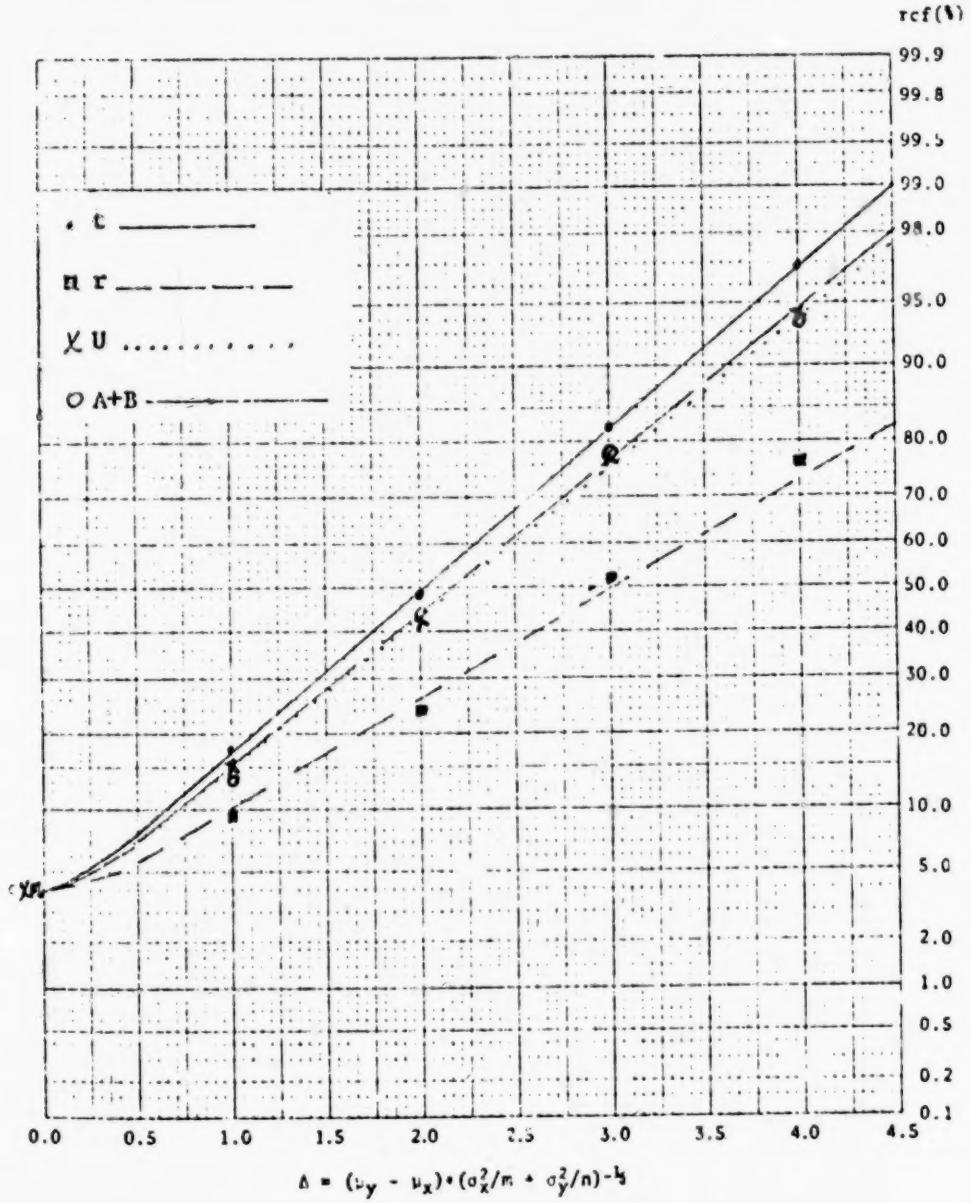


FIGURE 5

Empirical Power Values and Smoothed Power Curves
for t, r, and A+B for Uniform Distributions
m = n = 20, α = .05

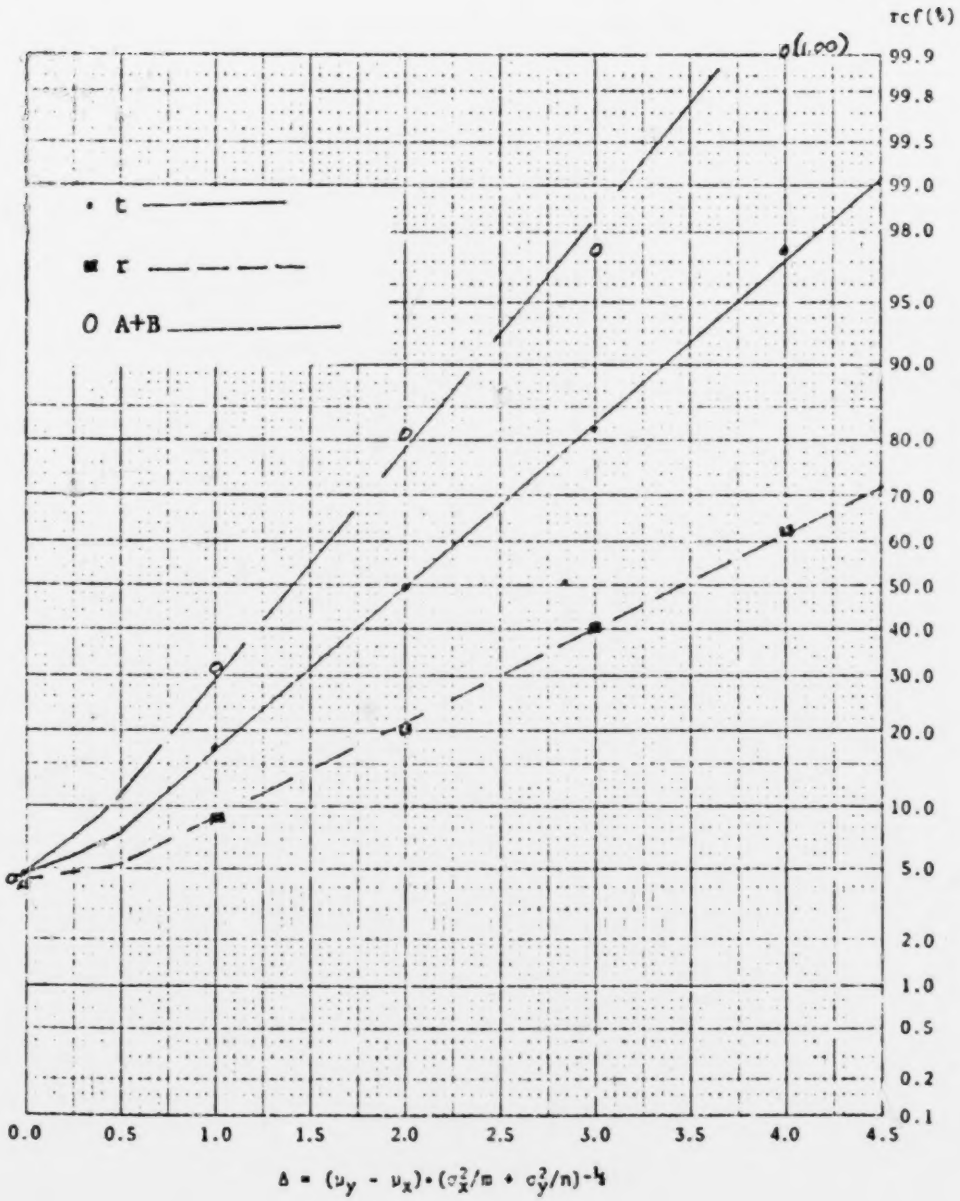
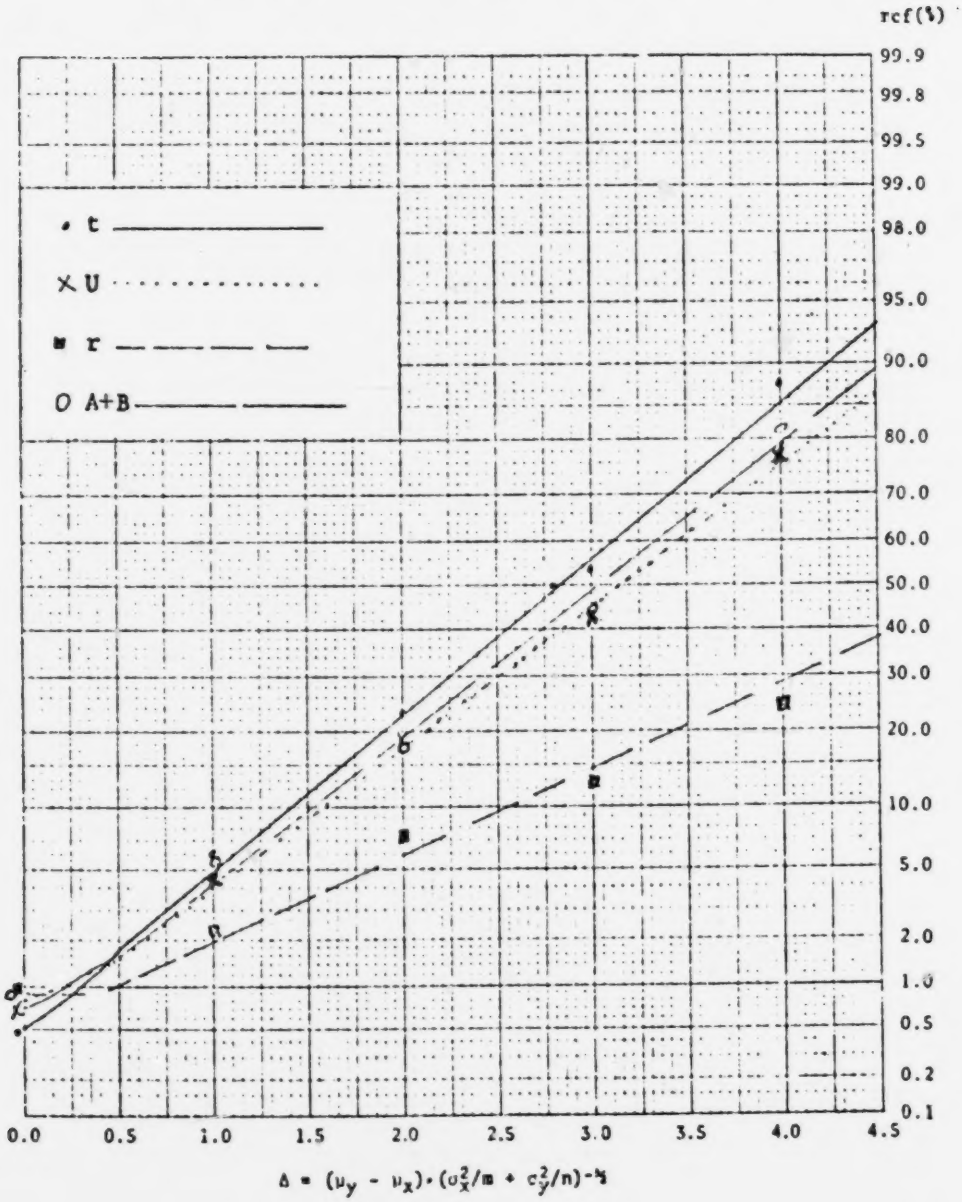


FIGURE 6

Empirical Power Values and Smoothed Power Curves
for t, U, r, and A+B for Uniform Distributions
m = 20, n = 5, α = .01



BIBLIOGRAPHY

Brandenburg, D. (1972). The Use of Multiple Matrix Sampling in Approximating an Entire Empirical Norms Distribution; Unpublished Ph. D. dissertation, University of Iowa.

Hajek, J. (1969). Non-Parametric Statistics; San Francisco, California; Holden-Day, Inc.

Hatch, L. O. and Posten, H. O. Robustness of the student-procedure: a survey. Research report No. 24, Department of Statistics, The University of Connecticut, 1966.

Pratoomraj, S. (1970). The Effect of Unequal Sample Sizes and Variance Heterogeneity and Non-Normality on Some Two-Sample Tests: An Empirical Investigation, Unpublished Ph. D. dissertation, University of Iowa.

Ramberg, J. and Schmeiser, B. (1972) An Approximate Method for Generating Symmetric Random Variables; Communications of the Association for Computing Machinery, 15, 987-990.

Tukey, J. W. (1959) A Quick, Compact, Two-Sample Test to Duckworth's Specifications, Technometrics, 1, 31-48.